Title
Regulatory remodeling in the allo-tetraploid frog Xenopus laevis.

Journal
Genome biology, 18(1)

ISSN
1474-7596

Authors
Elurbe, Dei M
Paranjpe, Sarita S
Georgiou, Georgios
et al.

Publication Date
2017-10-01

DOI
10.1186/s13059-017-1335-7

Peer reviewed

Elurbe et al. Genome Biology (2017) 18:198
DOI 10.1186/s13059-017-1335-7

Genome Biology

RESEARCH Open Access

Regulatory remodeling in the allo-tetraploid frog *Xenopus laevis*

Dei M. Elurbe[1†], Sarita S. Paranjpe[2†], Georgios Georgiou[2†], Ila van Kruijsbergen[2], Ozren Bogdanovic[3,4,5], Romain Gibeaux[6], Rebecca Heald[6], Ryan Lister[7], Martijn A. Huynen[1*], Simon J. van Heeringen[2*] and Gert Jan C. Veenstra[2*]

Abstract

Background: Genome duplication has played a pivotal role in the evolution of many eukaryotic lineages, including the vertebrates. A relatively recent vertebrate genome duplication is that in *Xenopus laevis*, which resulted from the hybridization of two closely related species about 17 million years ago. However, little is known about the consequences of this duplication at the level of the genome, the epigenome, and gene expression.

Results: The *X. laevis* genome consists of two subgenomes, referred to as L (long chromosomes) and S (short chromosomes), that originated from distinct diploid progenitors. Of the parental subgenomes, S chromosomes have degraded faster than L chromosomes from the point of genome duplication until the present day. Deletions appear to have the largest effect on pseudogene formation and loss of regulatory regions. Deleted regions are enriched for long DNA repeats and the flanking regions have high alignment scores, suggesting that non-allelic homologous recombination has played a significant role in the loss of DNA. To assess innovations in the *X. laevis* subgenomes we examined p300-bound enhancer peaks that are unique to one subgenome and absent from *X. tropicalis*. A large majority of new enhancers comprise transposable elements. Finally, to dissect early and late events following interspecific hybridization, we examined the epigenome and the enhancer landscape in *X. tropicalis* × *X. laevis* hybrid embryos. Strikingly, young *X. tropicalis* DNA transposons are derepressed and recruit p300 in hybrid embryos.

Conclusions: The results show that erosion of *X. laevis* genes and functional regulatory elements is associated with repeats and non-allelic homologous recombination and furthermore that young repeats have also contributed to the p300-bound regulatory landscape following hybridization and whole-genome duplication.

Keywords: Whole genome duplication, Interspecific hybridization, Genome evolution, Pseudogenes, Epigenomics, Enhancers

Background

Genome duplication is a major force in genome evolution that not only doubles the genetic material but also facilitates morphological innovations. In plants, whole-genome duplications (WGD) appear to occur more often than in animals [1] and some phenotypic innovations, like the origin of flowers, have been attributed to this phenomenon [2]. In animals, two rounds of WGD at the root of the vertebrate tree (~500 million years ago [Mya]) gave rise to the four HOX clusters and have led to the expansion of the neural synapse proteome [3]. It is likely that this facilitated an increase in the morphological complexity [4] and allowed an increase in the complexity in the vertebrate behavioral repertoire [5]. More recent genome duplications have been documented in fish, at the root of the teleost fish 320 Mya and in the common ancestor of salmonids 80 Mya [6]. Amphibians in general appear to have undergone many

* Correspondence: huynen@cmbi.ru.nl; s.vanheeringen@science.ru.nl; g.veenstra@science.ru.nl
†Equal contributors
[1]Radboud University Medical Center, Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands
[2]Radboud University, Faculty of Science, Department of Molecular Developmental Biology, Radboud Institute for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands
Full list of author information is available at the end of the article

polyploidizations, with natural polyploids in 15 Anuran and in four Urodelan families. In *Xenopus* (African clawed frogs), duplications have occurred on multiple occasions, giving rise to tetraploid, octoploid, and dodecaploid species [7]. One such duplication occurred in the ancestor of the amphibian *Xenopus laevis* 17 Mya [8]. The allo-tetraploid genome of *X. laevis* consists of two subgenomes, referred to as L (long chromosomes) and S (short chromosomes), that originated from distinct diploid progenitors [8]. Most of the additional genes that result from WGD events tend to be lost in evolution. In the case of allopolyploidy, this loss is biased to one of the parental subgenomes [9], a phenomenon referred to as biased fractionation. One explanation for biased fractionation is the variation in the level of gene expression between the homeologous chromosomes [10], with the lowest expressed gene having the highest probability of being lost because it would contribute less to fitness.

The effects of polyploidization on the epigenome have mainly been studied in plants, where correlations between the gene expression and epigenetic modifications have been observed between homeologous genes [11], but are not well characterized in animals. The epigenetic modifications found in chromatin (DNA methylation and post-translational modifications of histones) are involved in gene regulation during development and differentiation [12, 13]. A high density of methylated CpG dinucleotides is repressive towards transcription; conversely, the DNA of a large fraction of promoters is unmethylated. In addition, histone H3 in promoter-associated nucleosomes is tri-methylated on lysine 4 (H3K4me3) when the promoter is active. Active enhancers on the other hand are decorated with mono-methylated H3K4 (H3K4me1) and they also recruit the p300 (Ep300) co-activator which can acetylate histones. When genes are expressed, they not only recruit RNA polymerase II (RNAPII), responsible for the production of the messenger RNA, but the gene body will be decorated with H3K36me3, which is left in the wake of elongating RNAPII. Therefore, deep sequencing approaches to determine these biochemical properties in a given tissue or developmental stage can be used to interrogate the activity of genomic elements. This is highly relevant in the context of genomic evolution, as changes in gene expression caused by mutations in *cis*-regulatory elements are a major source of morphological change during evolution [14].

Here we ask how genome evolution and the epigenetic control of gene expression are related to interspecific hybridization and WGD. We compare functional regulatory elements in the L and S subgenomes of *X. laevis* embryos by chromatin immunoprecipitation (ChIP)-sequencing (ChIP-seq) of histone modifications, RNA-sequencing (RNA-seq), and whole-genome bisulfite sequencing (WGBS) and use *X. tropicalis*, a closely related diploid species, as a reference. We quantify the loss and the gain of genetic material and analyze how it has affected genes and gene-regulatory regions. Although genome evolution after the hybridization appears dominated by sequence loss, we also find evidence for the gain of functional elements. We specifically identify new subgenome-specific regulatory elements that recruit p300 and show that these are enriched for transposable elements (TEs). Finally, to assess the early gene-regulatory effects of hybridization, we analyze experimental interspecific *X. tropicalis* × *X. laevis* hybrids and we observe hybrid-specific p300 recruitment to DNA transposons, further highlighting the role of such elements in the evolution of gene regulation.

Results

The *X. laevis* L and S subgenomes show a bias in chromatin state and gene expression

To study the evolution of gene regulation in the context of WGD, we generated transcriptomic and epigenomic profiles in *X. laevis* early gastrula embryos (Nieuwkoop-Faber stage 10.5; Additional file 1). We performed RNA-seq and obtained epigenomic profiles using ChIP followed by deep sequencing (ChIP-seq). We generated ChIP-seq profiles for H3K4me3, associated with promoters of active genes, H3K36me3, associated with actively transcribed genes, the Polr2a subunit of RNA Polymerase II (RNAPII), and the transcription coactivator p300. In addition, we performed WGBS to obtain DNA methylation profiles [15]. The sequencing results and details are summarized in Additional file 1.

We created whole-genome alignments (see "Methods") to establish a framework for analysis of the epigenetic modifications in the two *X. laevis* subgenomes and in the *X. tropicalis* genome. Of the *X. laevis* L and S non-repetitive sequence, 61% and 59%, respectively, can be aligned with the orthologous *X. tropicalis* sequence. This allows for comparisons of the activity of genes and regulatory elements between homeologous regions. Figure 1 shows a region on *X. tropicalis* chromosome 8 containing four genes, together with the corresponding aligning sequences on chr8L and chr8S in *X. laevis*. The epigenomic profiles (H3K4me3, p300, RNAPII, and H3K36me3) of both *X. laevis* and *X. tropicalis* [16] are shown and the sequence conservation obtained from the whole gene alignment is illustrated by gray lines in the center of the plot. Regions that are conserved at both the sequence level and at the functional level (as measured by ChIP-seq) are highlighted. The *anp32e* gene is an example of a conserved gene that is expressed from all three genomes, as evidenced by H3K4me3 at the promoter and H3K36me3 and elongating RNAPII in the gene body. In contrast, expression of the *plekho1* gene has been lost from S. The gene is still present, but it is not active. There is no evidence of expression and

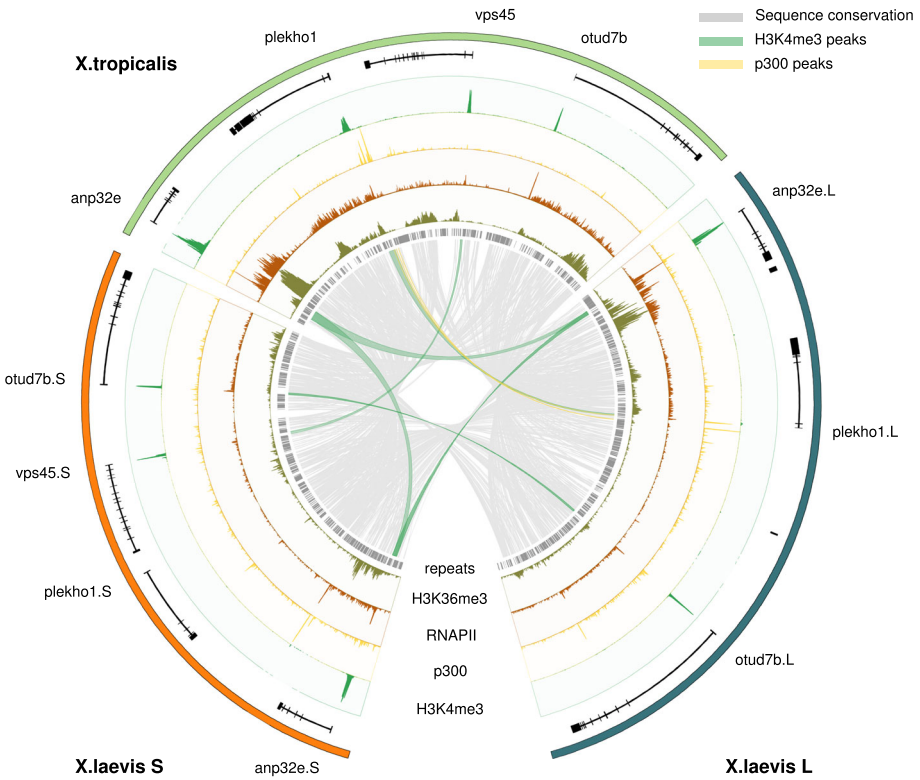

Fig. 1 Alignment of a region on chromosome 8 in *X. tropicalis* and the *X. laevis* L and S subgenomes annotated with experimental ChIP-seq data (gastrula-stage embryos; NF stage 10.5). Shown are the gene annotation (*black*), repeats (*gray*), ChIP-seq profiles for H3K4me3 (*green*), p300 (*yellow*), RNA Polymerase II (RNAPII; *brown*), and H3K36me3 (*dark green*). The sequence conservation is indicated by *gray lines*. Conserved H3K4me3 and p300 peaks are denoted by *green* and *yellow lines*, respectively. The anp32e gene is expressed in *X. tropicalis* and both the L and S subgenome of *X. laevis*. The *plekho1* gene, on the other hand, has lost promoter and enhancer activity on the *X. laevis* S locus and shows no experimental evidence of being expressed

both the H3K4me3 and the p300 signal are lost. Finally, the *vps45* gene is an example of a gene that is completely lost from L.

Next, we quantified gene expression patterns in the *X. laevis* subgenomes. Of the 17,303 genes expressed at stage 10.5, 9230 can be assigned to the L subgenome and 6685 to S. Of those expressed genes, 4972 are singletons located on L and 2646 on S. As reported previously [8], when both genes of a homeologous pair have detectable expression (3545 genes), the expression level is correlated (Pearson R = 0.60, $p < 1e\text{-}300$; Fig. 2a) and a minor but significant expression bias is detected (median expression difference of L compared to S = 5.7%; $p < 1e\text{-}4$; Wilcoxon signed-rank test). However, for many homeologs the expression bias is quite high, such that for one copy hardly any expression can be detected. Such non-expressed homeologs are located on both L and S, but occur more frequently on S (L: 494, S: 713; $p = 6.0e\text{-}11$, Fisher's exact test).

We examined whether the expression differences between the L and S homeologs could be explained by differential transcription regulation. We used the epigenomic profiles to assay the promoter state (H3K4me3, DNA methylation), enhancer activity (p300), and active expression (RNAPII, H3K36me3). The L subgenome has 38% more annotated genes than the S subgenome [8]. We observe the same trend for the regulatory elements. The number of H3K4me3 peaks, DNA-methylation free regions (see "Methods"), and p300 peaks is higher on L (28, 23, and 35%, respectively; Additional file 2). The overall effect is that there is no significant difference between the numbers of regulatory elements per gene for the two subgenomes.

To analyze the conservation of regulatory elements, we compared the H3K4me3 and p300 data to similar ChIP-seq profiles from *X. tropicalis* obtained at the equivalent developmental stage [16]. In general promoters are much more conserved than enhancers (Fig. 2b). From all H3K4me3 peaks in *X. tropicalis,* ~ 40% are conserved in *X. laevis*, while for the p300 peaks the conservation is only ~ 13% ($p < 1e\text{-}4$; Chi-squared test). This is congruent with the finding in mammals that enhancers evolve much more rapidly than promoters [17]. Whereas the number of conserved regulatory elements is lower in S than in L, the elements that can be aligned differ relatively little

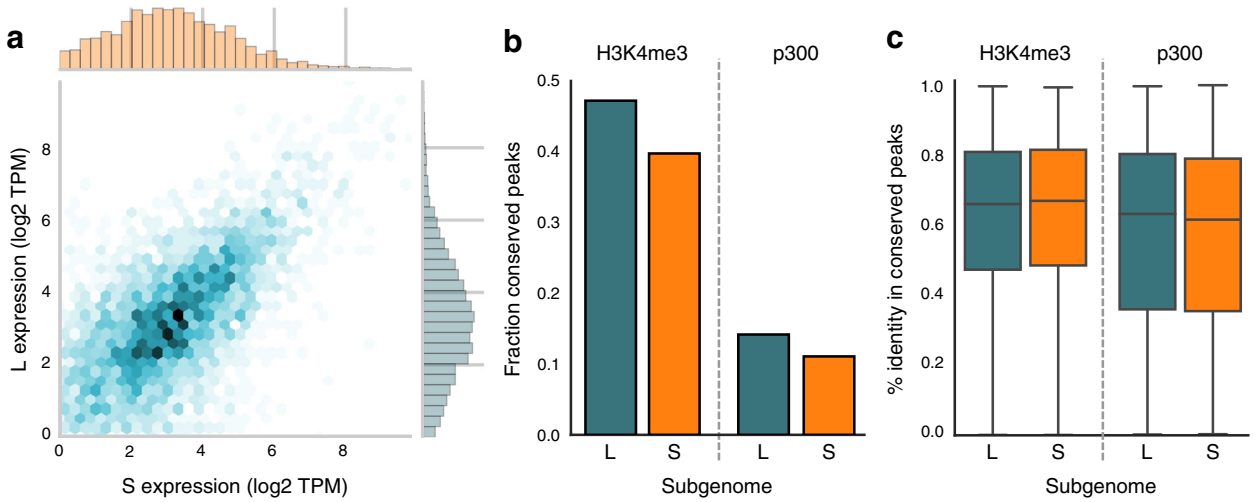

Fig. 2 **a** *Scatterplot* of the expression level (log2 TPM) of L and S homeologs that are both expressed. The expression level of homeolog genes is generally similar (Pearson R = 0.60, $p < 1e\text{-}300$). **b** Fraction of epigenetic signals ("peaks") conserved in *X. laevis* compared to *X. tropicalis*. Promoters appear more conserved than enhancers; S has lost more epigenetic elements than L. **c** Active functional elements are equally conserved between L and S compared to *X. tropicalis*. The background level of sequence conservation in fourfold degenerate sites from coding sequences with respect to *X. tropicalis* is 78.4% in L and 77.7% in S

at the sequence level and show over ~ 60% sequence identity (Fig. 2c).

These analyses show that the L and S subgenomes have evolved differently with respect to gene content [8] and regulatory elements. Many more genes from S are lower expressed than their homeologs in L than vice versa. The number of functional regulatory elements, as identified by H3K4me3 and p300 ChIP-seq, is proportional to a more profound loss of homeologous genes from the S subgenome. Next, we set out to determine the origin of this differential loss.

Large deletions are prominent in the S subgenome

The chromosomes of the *X. laevis* S subgenome are substantially shorter than the L chromosomes. The average size difference is 17.3% based on the assembled sequence [8] and 13.2% based on the karyotype [18]. To investigate the cause of these differences, we analyzed the pattern of deletions on both subgenomes. We called deleted regions based on the absence of conservation between the *X. laevis* subgenomes if they were at least partly conserved between one *X. laevis* subgenome and *X. tropicalis*. In addition, to be able to measure the size of the deletions, we required that the putative deleted regions were flanked on both sides by conserved sequences on both *X. laevis* subgenomes (Additional file 3: Figure S1). This resulted in a set of 19,109 deletions, of which 13,066 (68%) were deleted from S (LΔS) and 6043 (32%) were deleted from L (SΔL). There is a clear deletion bias towards S, which increases with the size of the deletion (Fig. 3a). These deletions affect genes and their regulatory sequences, as for example in the *glrx2* locus where the promoter and most of the exons have been lost from the S subgenome (Fig. 3b). We asked to what extent functional sequences in the L and S subgenomes are preserved (i.e. subject to fewer deletions) relative to the subgenome-specific deletion rates. To do that, we randomly redistributed the deletions per chromosome and compared the effect on various annotated and experimentally derived features. As we cannot assess these features before their deletion, we used the annotation and experimental data of the homeologous feature from the other subgenome as a proxy for the state in the genome from which that feature was deleted. The fold difference between the observed number of deleted basepairs and the expected number (mean of 1000 randomizations) is visualized in Fig. 3c. As expected, the frequency of deletions in intergenic regions and introns is similar relative to a uniform chromosomal distribution of deletions. The observed loss of exons on L is significantly lower than this randomized distribution ($p = 1.8e\text{-}20$; Fig. 3c). The fraction of exonic sequence that has disappeared is approximately fourfold less than intronic or intergenic sequence (Additional file 3: Figure S2). This is likely the result of negative selection against loss. By contrast, for subgenome S the fraction of exonic sequence that has been deleted is similar to the rest of S (Fig. 3c) and exonic sequences in S appear not to be under selection against deletion. To obtain more direct evidence of functional sequences, we examined the loss of genomic elements that are decorated with RNAPII and the active transcription histone mark H3K36me3 (IntronicTx, ExonicTx, see "Methods"), with the enhancer coactivator p300, or with the active promoter mark H3K4me3. There

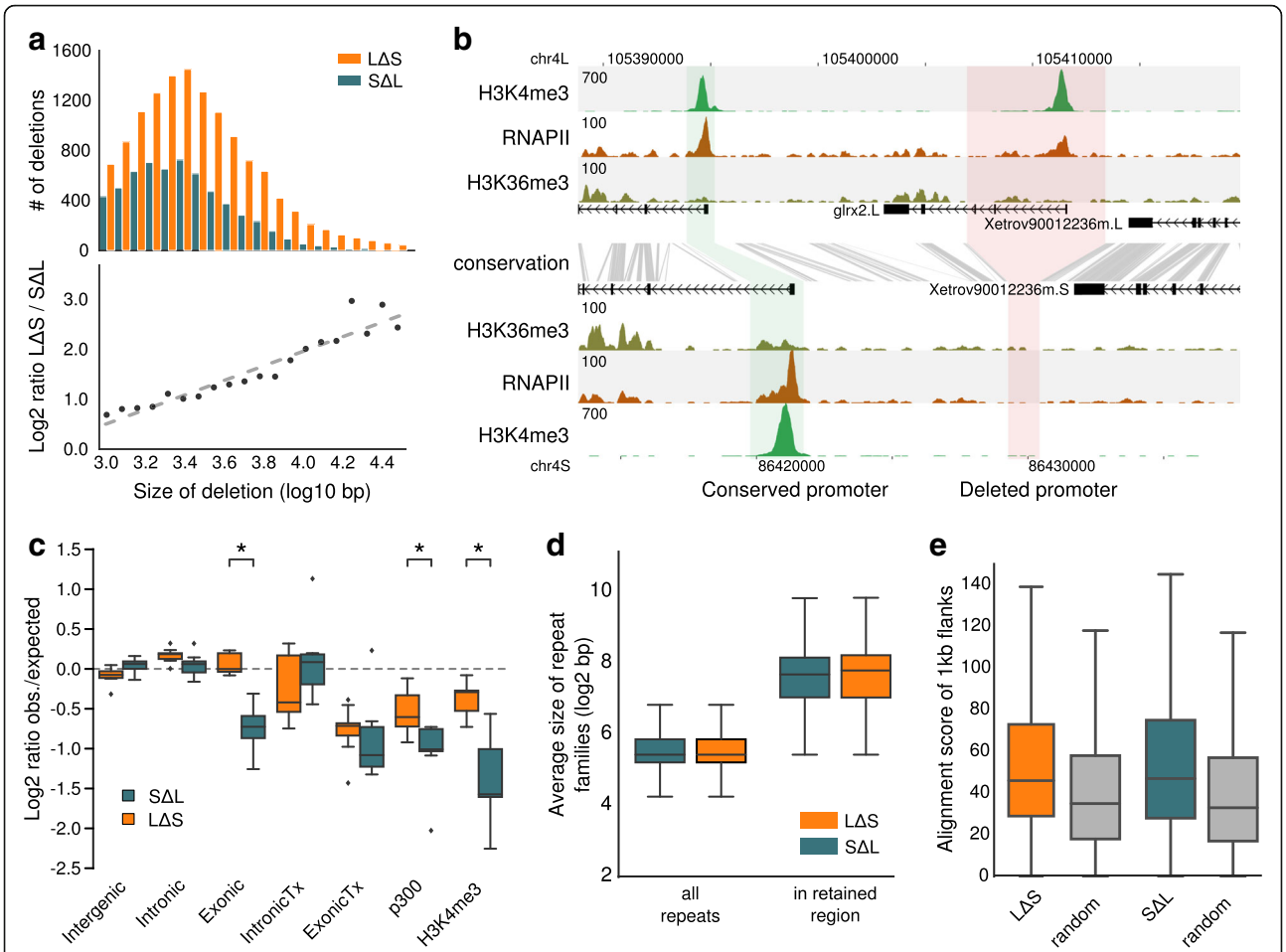

Fig. 3 The S subgenome has more and larger deletions than L. **a** Size frequency distribution of deletions (*top*) and size ratio of LΔS deletions relative to SΔL deletions as a function of deletion size (*bottom*). **b** An example of a gene (*grlx2*) that has lost the promoter on the S genome due to a deletion. Shown are the gene annotation (*black*), ChIP-seq profiles for H3K4me3 (*green*), RNAPII (*brown*), and H3K36me3 (*dark green*). The sequence conservation is indicated by *gray lines*. **c** The log2 fold difference between the observed number of deleted basepairs and the expected number (mean of 1000 randomizations). The fold difference is calculated per chromosome and summarized in a *boxplot*. *Intergenic* 1 kb distance from a gene, *Intronic* introns, *Exonic* UTRs + CDS, *IntronicTx* introns from genes actively transcribed, *ExonicTx* exons from genes actively transcribed, *p300* genomic fragments having a p300 peak, *H3K4me3* genomic fragments having a H3K4me3 peak. The *asterisks* mark significant differences between the L and S chromosomes ($p < 0.001$, Mann–Whitney U test). **d** Retained regions associated with deletions are enriched for relatively long repeats ($p <$ 1e-52 for both LΔS and SΔL; Mann–Whitney U test). **e** 1 kb flanks of the retained regions are more similar to each other than random genomic regions of the same size ($p <$ 1e-114 and 1e-83 for LΔS and SΔL, respectively; Mann–Whitney U test)

appears to be strong selection on both S and L against deletion of actively transcribed exons (Fig. 3c, middle panel; $p = 2.4e\text{-}4$ and $p = 2.3e\text{-}7$, respectively) but not of transcribed introns. Furthermore, active enhancers and promoters in S and in L have significantly fewer deletions compared to the uniform chromosomal distribution (Fig. 3c; $p = 8.4e\text{-}7$, $p = 8.4e\text{-}8$, $p = 1.4e\text{-}5$, and $p = 2.9e\text{-}12$, respectively) and therewith appear to be under selection against loss. There is a large difference in the number of deletions between L and S (Fig. 3a); however, this in itself is not necessarily the result of selection as it mostly affects non-functional sequences (Fig. S2a). We asked if, on top of this difference in absolute number, there is evidence for more selection against deletions in L than in S. We therefore compared the reduction in the loss of transcribed exons, promoters, and p300 elements relative to background loss between L and S. For all three the reduction in L appears to be larger than in S (Fig. 3c). For p300-bound enhancers and for H3K4me3-decorated promoters this difference in the reduction between L and S is significant ($p = 0.003$ and $p = 0.001$, respectively). This suggests that, aside from a higher deletion rate in S, there is also less selection against deletion of functional genetic elements in S than in L.

One of the possible sources of the loss of genomic DNA in the L and S subgenomes is non-allelic homologous recombination (NAHR), which is known to occur

between long repetitive elements on the same chromosome [19]. To test whether this phenomenon could be responsible for the genomic losses detected, we examined the length distribution of repetitive elements in retained regions, i.e. the homeologous regions of the sequences that were lost in one of the subgenomes (Fig. 3d). Indeed, we observe that repetitive elements are on average 3.7 times longer ($p < 1e\text{-}52$; Mann–Whitney U test) compared to random genomic sequences (Fig. 3d). Furthermore, the flanks of the retained regions (L for LΔS and S for SΔL, respectively) tend to be more similar to each other than random genomic sequences ($p < 1e\text{-}83$; Mann–Whitney U test; Fig. 3e). Nevertheless, the current density of repetitive elements is similar in the L and S subgenomes (Additional file 3: Figure S3), indicating that repeat density alone does not cause biased sequence loss on S chromosomes. These observations suggest that NAHR of ancient repeats has played a significant role in the deletions of regions from both subgenomes; the overall sequence loss is much more prevalent on the S chromosomes (Fig. 3a). To estimate when in the evolution these deletions and other types of mutations occurred, we dated the origin of the pseudogenes that they caused.

High levels of pseudogenization started after hybridization and continue to the present

To date the pseudogenes, we aligned them with the protein-coding regions in L, S, and the outgroup *X. tropicalis* (see the "Search and alignment of orthologs and evolution rates" section in "Methods"). The coding regions in S are generally less conserved than in L, especially regarding synonymous substitutions (Ks, Fig. 4a, $p < 2.2e\text{-}16$; Wilcoxon signed-rank test). However, the ratio between non-synonymous and synonymous substitutions (Ka/Ks) is only slightly higher in S compared to L (Fig. 4b, $p < 2.2e\text{-}16$; Wilcoxon signed-rank test). The difference in Ks between the L and S subgenomes shows that S has been subject to moderately higher mutation rates than L. In order to examine whether the relatively high level of mutations in the S genome persists to this day, we examined the level of SNPs separating the published inbred genome [8] and the progeny of two outbred individuals (see the "SNP calling" section in "Methods"). We observe that the level of SNPs in the S genome is 3% higher than in the L genome in intergenic ($p = 5e\text{-}136$; Chi-squared test) and intronic regions ($p = 8e\text{-}101$; Chi-squared test). A similar difference is observed in fourfold degenerate (4D) positions of coding DNA (also assumed to be under relaxed constraint) but this is not statistically significant (Additional file 4). The 4D positions exhibit a SNP density higher than in non-coding DNA; this correlates with an overrepresentation of CpGs in coding DNA (Additional file 3: Figure S4) and has been observed before in human genomes [20].

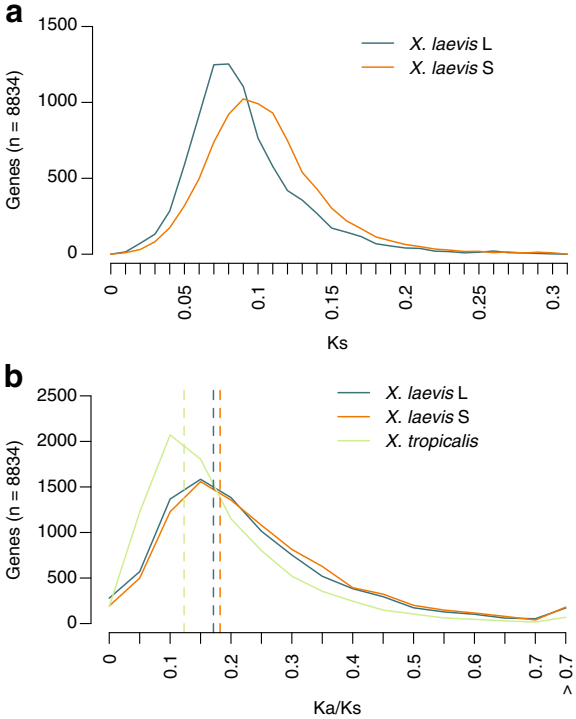

Fig. 4 The S subgenome has a higher mutation rate than L. Only genes which none of the L or S copies fall into the pseudogene category are considered. **a** Ks distribution per subgenome in *X. laevis*. **b** Ka/Ks in *X. laevis* and *X. tropicalis*

Given that the hybridization event occurred 17 Mya [8], the higher SNP density in S relative to L (Additional file 4) cannot be a relic from the time before the hybridization (Additional file 5) and it suggests that the relatively high rate of genome degradation in S continues to this day. To examine the continuity of this genome degradation, we dated unitary pseudogenes [21] caused by point mutations and/or deletion-related events (Fig. 5a). We distinguish four, non-exclusive types of pseudogenes: genes that contain a premature stop codon; genes of which the coding sequence is at least 50% shorter than their homeolog and their ortholog in *X. tropicalis*; genes that have lost at least the 75% of their promoter relative to their homeologs that do have a promoter decorated with H3K4me3 in embryos; and genes that contain a frameshift. We furthermore required for each class that the pseudogene candidate is expressed at least tenfold lower than its homeolog. In all cases, we do observe that the rate of pseudogenization has increased dramatically around 18 Mya, i.e. close to the inferred date of the hybridization, and that that rate is ~ 2.3-fold higher in S than in L (Fig. 5a). Furthermore, this rate continues to be high until this day for every class considered (Fig. 5b). We obtained

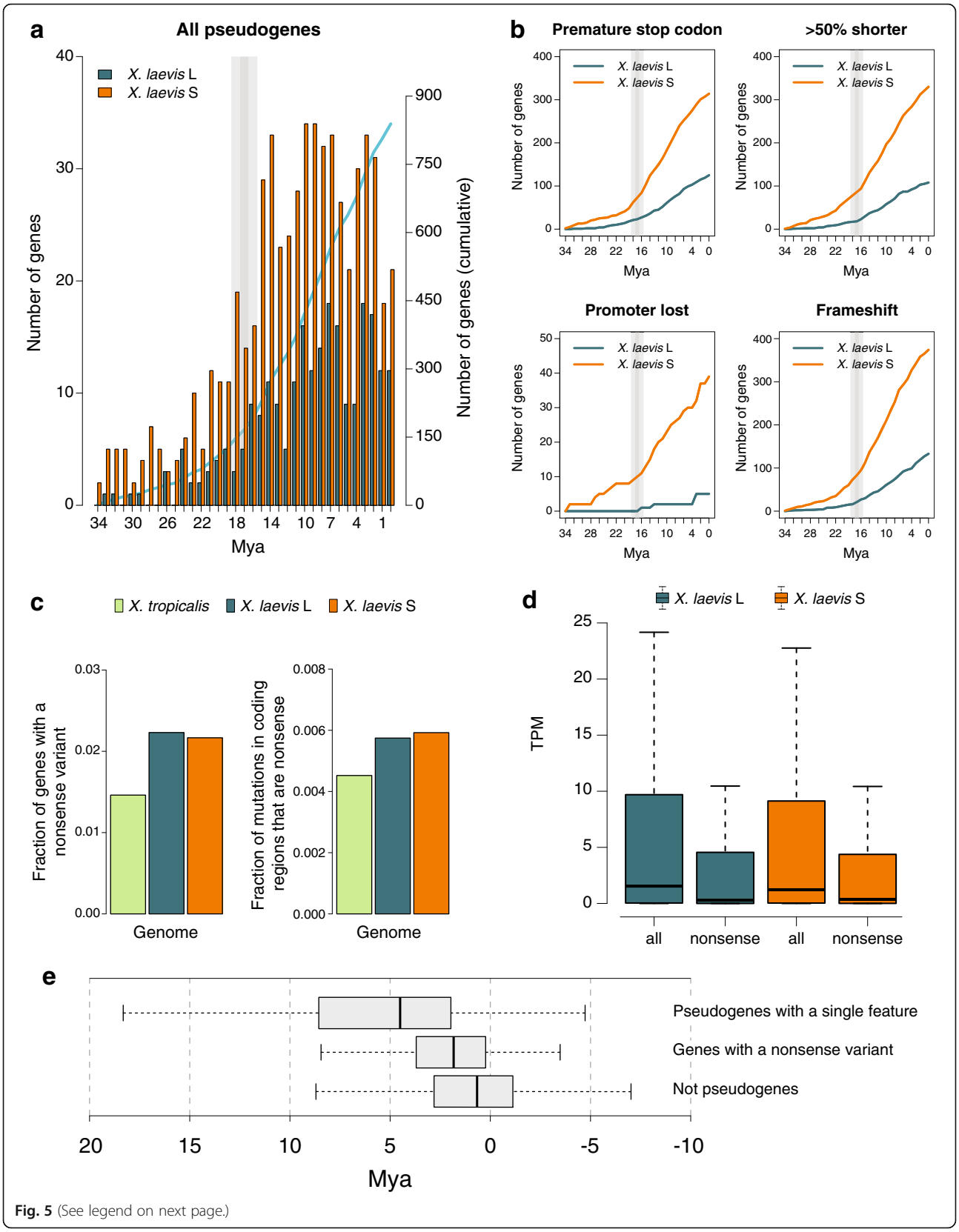

Fig. 5 (See legend on next page.)

(See figure on previous page.)
Fig. 5 Pseudogenization rate has increased after hybridization. **a** Number of likely pseudogenes (i.e. genes having one or more pseudogene feature and no expression while their homeolog is expressed) binned by predicted date of pseudogenization event. **b** Pseudogenes with different (non-exclusive) pseudogene features and their sum over the years. **c** *Left*: fraction of genes that have a nonsense variant in the population. *Right*: fraction of mutations in coding regions that introduce a premature stop codon. **d** Expression of genes with and without a nonsense variant present in the population. **e** Distribution of predicted pseudogenization time (including one-to-one orthologs of human, mouse, and chicken) for genes with a single pseudogene feature and a tenfold lower expression than the homeolog (*top*), for genes with a nonsense variant present in the population of *X. laevis* (*middle*) and for genes that do not present any feature for pseudogenization and whose expression is less than twofold different between homeologs (*bottom*)

very similar results when we included one-to-one orthologs from additional species in the dating of the pseudogenes and bootstrapped the results per gene to obtain confidence intervals (see the "Bootstrapping pseudogene dates" section in "Methods") (Additional file 3: Figure S5). When we separate the pseudogenes into non-overlapping classes we observe that deletions are a prevalent cause of pseudogenization (39% and 44% on L and S, respectively) and, as expected, the older pseudogenes are affected by more than one type of damage (Additional file 3: Figure S6). Pseudogenization after genome duplication has been observed to affect certain classes of protein functions more than others, with metabolic functions often being the first ones to be lost relative to regulatory proteins [6]. Indeed, when we date the loss of genes in the function categories associated with the loss, we find an overrepresentation of various metabolic processes, with the pseudogenes belonging to those categories dating often shortly after the WGD event (Additional file 3: Figure S7). We found no evidence for the preferential loss of complete complexes rather than partial complexes, e.g. for dimers the fraction of cases where of both genes only a single copy was left (17.6%), was not higher than the expected percentage if we assumed the losses of the genes from complexes to be independent from each other (18.0%) (see "Methods"). To test for the influence of a potential dosage effect on gene loss, we compared the predicted genome-wide haploinsufficiency score (GHIS) [22] of the human ortholog of *X. laevis* homoeolog and singleton genes (Additional file 3: Figure S8). Singletons indeed have a significantly lower GHIS score than homeologs ($p = 1.1e\text{-}17$; Mann–Whitney U test), although the difference is minor (3.0%).

To find independent evidence that the rate of pseudogenization in *X. laevis* remains high until the present, we examined genes that appeared to be polymorphic with respect to their pseudogene state, i.e. we searched for protein truncating variants (PTVs) (variants which potentially disrupt protein-coding genes) in the progeny of two of our outbred genomes (see the "SNP calling" section in "Methods") relative to the published inbred genome [8]. Among all possible PTVs, we limited the analysis to SNPs that introduce a premature stop codon (nonsense mutations), as they can be called relatively reliably [23]. As a reference, we compared the nonsense SNP density with the one we measured in *X. tropicalis* using the same type of data and settings to call the SNPs, i.e. the progeny of two outbred genomes. In the 23,667 annotated genes in L and 16,939 in S, we detect 528 (2.23%) and 367 (2.17%) genes with at least one loss of function (LOF) variant. In contrast, in the 26,550 genes of *X. tropicalis*, we detect only 388 (1.46%) LOF variants (Fig. 5c, left). When normalizing the nonsense variants by the total number of SNPs in coding regions per (sub) genome, the fraction of premature stop variants in S (5.9e-3) is slightly higher than that in L (5.7e-3) while both are substantially and significantly higher than in *X. tropicalis* (4.5e-3; $p < 0.001$ for both comparisons; Chi-squared test; Fig. 5c, right). To substantiate that the selected PTVs are indeed hallmarks of incipient pseudogenes, we compared their expression with the expression of the other genes in their respective (sub)genome and found that genes with a SNP introducing a premature stop codon have a significantly lower expression (Fig. 5d). Second, we used the equation for dating of unitary pseudogenes to estimate the time of loss of selection in the PTV containing genes. We found that genes with this type of variants present in the population show evidence of loss of selection when compared to the set of genes that are not pseudogenes ($p = 1e\text{-}5$; Student's *t*-test; Fig. 5e) and that this loss of selection is more recent than for pseudogenes with only a single feature for pseudogenization that is fixed in the population ($p = 5.6e\text{-}7$; Student's *t*-test; Fig. 5e). That we find a higher level of SNPs in S than in L cannot be a relic from the time before the hybridization in which the S species may have had a higher SNP density than L, given that the hybridization occurred 17 Mya (Supplemental note). Altogether, these results suggest that, in addition to deletions, a higher mutation rate and a more relaxed selection pressure in S has contributed to the differences that the subgenomes present nowadays, including differential gene loss. This gene loss continues to be at a higher rate than in a closely related diploid species.

Transposons have contributed subgenome-specific enhancer elements

The results described above document the pervasive loss and ongoing decay of coding and regulatory sequences after interspecific hybridization genome duplication. We

next asked to what extent regulatory innovations have contributed to genomic evolution of this species. At many loci, the profile of p300 recruitment is remarkably different between L and S loci, with differences in both p300 peak intensity and number of peak regions across homeologous loci, for example in the *slc2a2* locus (Fig. 6a). We identified 2451 subgenome-specific p300 peaks lacking any conservation with either the other subgenome or *X. tropicalis* (colloquially referred to as "new" enhancers). There are similar numbers of these non-conserved subgenome-specific p300-bound elements in the L subgenome (n = 1214) and the S subgenome (n = 1237).

Because new sequences can be acquired by transposition, we examined the overlap of subgenome-specific enhancers with annotated repeats and found that 87% (2143 of 2451; overlap > 50%) are associated with annotated repeats, compared to 24% (5557 of 23,017) of all enhancers ($p < 1e\text{-}308$; hypergeometric test). Three repeats (designated REM1, Kolobok-T2, and family-131) were particularly enriched; individually they overlap with 37–53% of the subgenome-specific p300 peaks, compared to 3–9% at other p300 peaks (Fig. 6b). Together these three annotations account for 1338 (54%) of new enhancers, 862 of which have all three annotations overlapping at the same location. They form a 650-bp sequence with an almost perfect 195-bp terminal inverted repeat (TIR), the most terminal 65 bp of which shows 83–90% similarity with the TIRs of a Kolobok-family DNA transposon present in *X. tropicalis* (Additional file 3: Figure S9). This specific Kolobok DNA transposon carries the REM1 interspersed repeat and is present almost exclusively in *X. laevis* (8833 and 8802 copies in L and S, respectively, vs. four copies in *X. tropicalis*), suggesting that it is a relatively young TE that proliferated after the split with *X. tropicalis*. It carries several transcription factor (TF) motifs, including the Eomes T-box motif and the Six3/Six6 homeobox motif (Fig. 6c).

We examined the correlation of the new Kolobok enhancers with gene expression and found that genes with a transcription start site within 5 kb of these subgenome-specific Kolobok enhancers are more highly expressed than other genes in that subgenome ($p = 1e\text{-}4$ for L and $p = 8e\text{-}5$; Mann–Whitney U test) (Additional file 3: Figure S10), suggesting that the new enhancers are inserted close to active genes and/or promote the expression of these genes.

Regulatory remodeling by transposons in *X. tropicalis* × *X. laevis* hybrids

The gene expression (Fig. 2) and p300 recruitment (Fig. 6) differences between the L and S subgenomes may have been caused by regulatory incompatibilities affecting enhancer activity or DNA methylation, which could act immediately upon interspecific hybridization. Alternatively, these differences may represent the long-term effects of genomic co-evolution of the two subgenomes. To examine whether the differences between the two subgenomes were caused by the hybridization event itself, we determined the immediate effect of hybridization on DNA methylation and the patterns of H3K4me3 and p300 enrichment at regulatory regions. We generated embryos obtained by fertilization of *X. laevis* eggs (LE) with *X. tropicalis sperm* (TS). The resulting LETS hybrid embryos were compared to normal *laevis* (LELS) and *tropicalis* (TETS) embryos. The reverse hybrid (TELS) was not viable, as previously described [24].

To examine the early potential changes in DNA methylation, we performed WGBS on the DNA of LETS, LELS, and TETS embryos. The overall methylation in hybrid and normal embryos is almost identical at 92%. We identified a total of 709 differentially methylated regions (DMR) (false discovery rate [FDR] = 0.05); 181 and 72 hypermethylated and 384 and 72 hypomethylated regions in respectively the *X. laevis* and *X. tropicalis* genomes. This reflects both gain and loss of DNA methylation in the subgenomes of LETS hybrid embryos (Fig. 7f, g). There is no evidence in the underlying DNA sequence signatures for these regions being related to gene-regulatory regions (Additional file 3: Figure S11a–d). They are also not in close proximity of genes and may represent regions with inherently unstable DNA methylation. The global pattern of H3K4 trimethylation at promoters is also quite similar in hybrids and normal embryos; less than ten peaks changed in hybrid embryos relative to normal embryos (Additional file 3: Figure S11e).

Recruitment sites of p300, however, are specifically gained and lost at several subsets of *X. tropicalis* genomic loci in hybrid embryos (Fig. 7a); 629 p300 recruitment sites were gained (a 2.6% increase relative to normal *X. tropicalis* embryos), whereas just 67 p300-bound regions were lost (adjusted p value cutoff 1e-5). In the *X. laevis* part of the hybrid genome, none were lost or gained (Fig. 7a), indicating that the changes in the hybrid are biased towards the paternal *tropicalis* genome. To assess the epigenetic state of the gained and lost p300-binding regions, we used our epigenome reference maps of histone modifications in *X. tropicalis* [16]. Among all the marks tested, only H3K9me3 was significantly enriched, specifically at sites of gained p300 recruitment (Fig. 7b), suggesting that these regions are heterochromatic in normal (TETS) embryos but can recruit the p300 co-activator in LETS hybrid embryos.

While examining the p300 hybrid-specific recruitment sites, we noticed that transposable elements were present at many locations (Fig. 7c, d); 82% of the hybrid-specific p300 peaks overlapped more than 50% with annotated repeats. We therefore examined the occurrence of specific repeats at gained p300 sites and found that three repeat annotations (family - 451, 203, and 189) were strongly

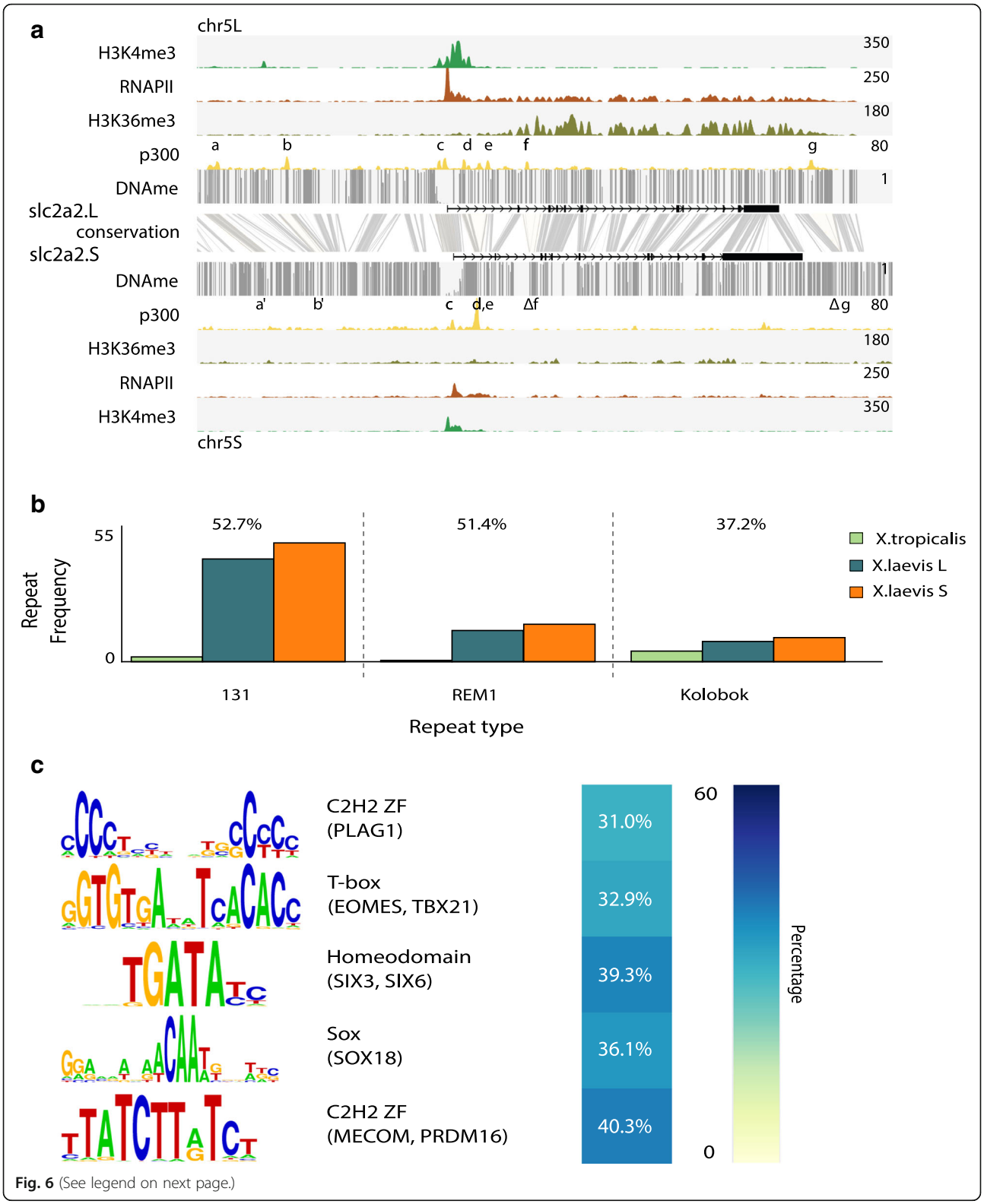

Fig. 6 (See legend on next page.)

(See figure on previous page.)
Fig. 6 Subgenome-specific recruitment of p300 is associated with TEs. Subgenome-specific p300 peaks are enriched for TEs carrying transcription factor (TF) motifs active in early development. **a** Differential regulation of the slc2a2 homeologs at stage 10.5. Shown are the genomic profiles of H3K4me3 (*green*), RNA Polymerase II (RNAPII; *purple*), H3K36me3 (*blue*), and p300 (*yellow*) ChIP-seq tracks, as well as DNA methylation levels determined by WGBS (*gray*). The *top panel* shows slc2a2.L, which is highly expressed, as evidenced by RNAPII and H3K36me3, and has a number of active enhancers (**a**–**g**), while slc2a2.S, shown in the *bottom panel*, is expressed at a lower rate. The conservation between the L and S genomic sequence is shown in *gray* between the panels. Differential enhancers between L and S are highlighted in *yellow*, which illustrates lost enhancer function (**a**, **b**), conserved enhancer function (**c**–**e**), and deleted enhancers (**f**, **g**). **b** Subgenome-specific p300 peaks are associated with DNA transposon repeats (threshold $p \leq$ 10e-4, twofold enrichment compared to all *X. laevis* peaks and present at least in 15% of the peaks). The *barplots* show the frequency of occurrence of each of the three repeat types per megabase in the three (sub)genomes. Over the *bars* is represented the percentage of subgenome-specific peaks overlapping with the corresponding repeat. **c** TF found to be enriched in the subgenome-specific p300 peaks (threshold $p \leq$ 10e-4, threefold enrichment compared to all *X. laevis* peaks and present at least in 20% of the peaks)

enriched (p = 1e-5; hypergeometric test), each accounting for 20–37% of all newly gained p300 peaks, whereas they only overlap with < 1% of other p300 peaks (Fig. 7c, lower panel). The three repeat annotations strongly co-occur and form a 1.3-kb sequence with a 200-bp imperfect TIR, which shows ~ 80% similarity with those of known PiggyBac-N2A DNA transposons (Additional file 3: Figure S12). We recently found that DNA transposons that are heterochromatinized by H3K9me3 in *X. tropicalis* embryos are relatively young relative to other TEs [25]. Indeed, the piggyBac DNA transposons that gain p300 binding in hybrids are much less abundant in *X. laevis* than in *X. tropicalis*, suggesting that these relatively young transposons get derepressed in the *X. laevis* egg which has had little prior exposure to this transposon. These elements also carry transcription factor binding sites. Nine motifs are enriched (p = 1e-5; hypergeometric test) and are present in 10–35% of gained p300 recruitment sites, compared to a 1–3% prevalence of these motifs in other p300 peaks (Fig. 7e). These DNA-binding motifs represent binding sites of Homeodomain and T-box binding factors, which are abundantly expressed during early embryogenesis.

These results document DNA transposon-associated p300 recruitment and DNA methylation instability in experimental interspecific hybrids.

Discussion

The genomes of the parental *Xenopus* species that gave rise to *X. laevis* through interspecific hybridization have remarkably been maintained as separate and recognizable subgenomes propagated on different sets of chromosomes [8]. These clearly distinguishable subgenomes allow detailed analyses of the patterns of (epi)genomic loss and regulatory remodeling.

The loss of genes, regulatory elements, and genomic sequence is caused predominantly by deletions and mutations in both subgenomes, which erode the S subgenome more strongly than the L subgenome. Such biased loss of genes has been observed in polyploid plant species and has been suggested to be a general result of allopolyploidization, in contrast to auto-polyploidies where the subgenomes are indistinguishable and degrade at a similar rate [9]. As to why one particular subgenome erodes more quickly than another, one hypothesis is that interspecific hybridization generates a crisis, referred to as "genomic shock," for example by transposon reactivation on one of the subgenomes which can disrupt coding sequences [26]. Consistent with this possibility is the proliferation of S-specific Mariner DNA transposons in *X. laevis* at the time of hybridization [8]. Also consistent with transposon reactivation are our results from artificial *X. tropicalis* × *X. laevis* hybrids (LETS, *X. laevis* eggs, *X. tropicalis* sperm), in which a set of *X. tropicalis*-specific DNA transposons recruits the p300 co-activator in the hybrid, whereas normally they are repressed by H3K9me3. Relatively young DNA transposons are heterochromatinized with H3K9me3 [25], but when introduced into eggs that have been little exposed to these transposons these mechanisms may fail. We have not been able to detect transposon expansion in the short time of *Xenopus* hybrid embryogenesis (data not shown), but together the observations suggest that transposon reactivation can contribute to genomic perturbations in hybrids. Similarly, in the Atlantic salmon, which has undergone several (320 Mya, 80 Mya) whole genome duplications, transposon expansion has been associated with the whole genome duplication event and with chromosome rearrangements [6].

In contrast to these short-term effects of hybridization, our analyses indicate that new pseudogenes continue to arise, both by mutations that cause premature stop codons, and by deletions that truncate the coding region or delete intergenic or promoter regulatory sequences. An elevated rate of pseudogene formation is observed on both the L and S subgenomes since the time of hybridization (~17 Mya, cf. Fig. 5) up to the present day, suggesting genome erosion is a continuous process that has been and still is higher on S compared to L. Consistent with this result is a mildly elevated level of SNPs observed in S relative to L (Fig. 4; Additional file 4). The cause of the higher mutation rate of the S subgenome is unknown. The local mutation rate has been shown to correlate with replication timing [27] and it is possible that there are subtle but consistent differences in replication timing between the two subgenomes. It can also be due to differences in

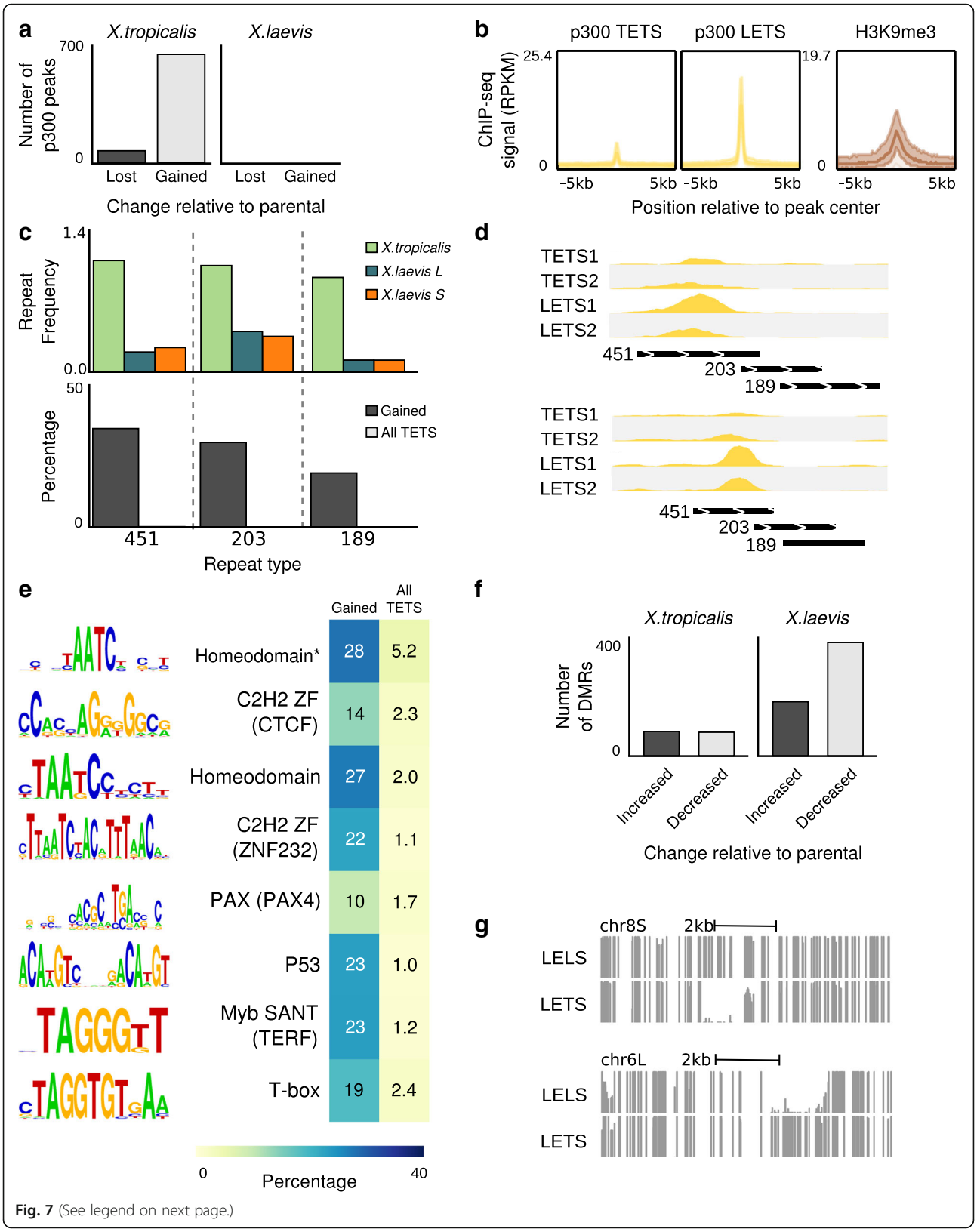

Fig. 7 (See legend on next page.)

(See figure on previous page.)
Fig. 7 a Changes in p300 recruitment in LETS hybrids. In the *X. tropicalis* genome there are new hybridization-induced peaks as well as peaks that disappeared after hybridization. In the *X. laevis* genome there are no changes. **b** Newly introduced peaks appear to be repressed by H3K9me3 in *X. tropicalis* embryos. **c** *Bottom*: a significant number of hybrid-specific peaks are associated with DNA transposon repeats (threshold $p \leq$ 10e-6, > 20 times fold enrichment compared to all *X. tropicalis* peaks and present at least in 10% of the peaks). *Top*: the *bar plots* show the frequency of occurrence of Motif:lcl|rnd-1_family-451_DNA, Motif:rnd-1_family-203 and Motif:lcl|rnd-1_family-189_DNA_PiggyBac repeats per megabase in the three (sub)genomes. Those repeats are *X. tropicalis*-specific, as they occur more often compared to *X. laevis* genomes. **d** Profiles of *X. tropicalis* embryos p300 and LETS hybrid p300 in *X. tropicalis* hybridization-induced peaks loci. New peaks overlap with DNA transposon repeats. **e** Newly introduced peaks found to be enriched in TF DNA binding sites (threshold $p \leq$ 10e-6, fivefold enrichment compared to all *X. tropicalis* peaks and present at least in 10% of the peaks). The TFs that can bind these motifs include Homeobox factors, C2H2 Zinc finger proteins (CTCF, ZNF232), PAX4, TERF, and T-box factors. The AATC motif, marked by an *asterisk*, is annotated in TRANSFAC as a GATA1 motif, but closely resembles a Paired Homeobox consensus motif. **f** DMRs in hybrid embryos. **g** DNA methylation profiles showing the DNA methylation instability in LETS hybrids

background selection [28], in which selection against non-neutral variants would also reduce neutral variation in their vicinity.

All in all, the higher level of genome degradation in S relative to L appears to be the result of a slightly higher mutation rate and a considerably higher deletion rate in S, combined with less selection against the loss of (epi)-genetic elements in S than in L. The higher deletion and mutation rates are supported by higher numbers of deletions and SNPs in regions that appear not to be under selection: intergenic regions; introns; and redundant coding positions. Reduced selection against the loss of genetic elements from S relative to L is supported by a larger difference in the loss of p300 peaks and promoters relative to the background in the L subgenome than in the S subgenome and a slightly but significantly lower Ka/Ks ratio in the L subgenome relative to the S subgenome.

The deletions bear the hallmarks of NAHR [29]; the retained regions in the other subgenome are enriched for ancient repeats and the sequence similarity between the flanks of the region is higher than expected by chance. The S chromosomes have also experienced significantly more rearrangements including inversions [8]. Normally, in meiotic recombination double strand breaks are fixed using allelic sequences. In the absence of proper chromosome pairing, other non-allelic homologous sequences, for example repeats in the same chromosome, are used for double-strand break repair, leading to deletions and inversions [29]. Interestingly, Prdm9, a fast-evolving mammalian DNA-binding protein involved in meiotic chromosome pairing and recombination hotspot selection, has been implicated in hybrid sterility in mouse [30, 31]. There is no known one-to-one ortholog of Prdm9 in *Xenopus* and the L and S subgenome-encoded proteins involved in meiotic double strand break repair are also not fully known, but it is conceivable that their skewed expression or activity is involved in subgenome-biased NAHR.

The results reported here identify a major role for repetitive elements in subgenome bias, gene loss, and regulatory remodeling. Not only is sequence loss by NAHR linked to repeats, subgenome-specific acquisition of enhancer elements is also overwhelmingly associated with TEs. Moreover, young transposons also gain p300 recruitment in *X. tropicalis* × *X. laevis* hybrids. DNA transposons can contribute sequence variation to the genome, which can affect gene expression by changing the local chromatin state at the site of insertion, resulting in metastable epi-alleles [26]. Once a host is invaded, TEs usually duplicate freely before they become repressed. When introduced in relatively unexposed eggs this repression may be lost. Interestingly, TEs can be co-opted as enhancers for the regulation of developmental genes [32, 33]. TFs have been found to bind to TEs with open and active chromatin signatures in both human and mouse cells, but the binding patterns were largely different between the two species [34], suggesting that transposons contribute to regulatory change during evolution. In addition to the potentially large and sudden changes in regulatory potential caused by transposition, mutational changes are known to cause TF-binding sites to be lost and gained [17, 35] causing turnover and change in the regulatory landscape over longer time scales.

Conclusions

It is not known exactly how the ancient two rounds of whole genome duplications at the root of the vertebrate tree have contributed to genome evolution. Its analysis is confounded by the pervasive loss of homeologs over hundreds of millions of years and the absence of tractable subgenomes. The *X. laevis* interspecific hybridization and genome duplication event is one of the most recent vertebrate genome duplications. Excitingly, the clearly distinguishable chromosomes of different parental origins allow for reconstruction of the parental genomes. We have found evidence for a pervasive influence of repetitive elements, driving gene loss, and genomic sequence loss through NAHR, in addition to remodeling of the regulatory landscape through transposon-mediated gain of coactivator recruitment. In combination with experimental interspecific hybrids, *Xenopus* can therefore be a powerful new model system to distinguish the short- and

long-term consequences of hybridization and to study the mechanisms of vertebrate genome evolution.

Methods

Animal procedures

Embryos were generated using in vitro fertilization (IVF) with outbred animals, including LELS embryos (laevis eggs–laevis sperm), TETS embryos (tropicalis eggs–tropicalis sperm), and LETS embryos (laevis eggs–tropicalis sperm). *X. laevis* female frogs were injected with 500 U of human chorionic gonadotropin (hCG, BREVACTID 1500 I.E) 16 h before IVF. A *X. laevis* male was sacrificed and isolated testis was macerated in 2 mL Marc's Modified Ringer's medium (MMR) to be used immediately for fertilization. Both male and female *X. tropicalis* frogs were primed with 100 and 15U of hCG 48 h before IVF. Five hours before egg laying, females were boosted with 150 U of hCG. Male testis was always isolated fresh. The testis was macerated in 2 mL FCS-L15 (10% fetal calf serum–90% L15 medium) cocktail and used immediately for IVF. LETS embryos were obtained similarly using species and sex-specific hormonal stimulation as described above. Once the macerated sperm suspension was mixed vigorously over the layered eggs, they were left undisturbed for three minutes and then the Petri dish was flooded with 25% MMR for the fertilized *X. laevis* eggs (LELS and LETS) and 10% MMR was added to the fertilized *X. tropicalis* eggs (TETS). Embryos were cultured at 25 °C. The jelly coats were removed 4 h post fertilization (hpf) using 2% cysteine in 25% MMR (pH 8.0) for LELS and LETS and using 3% cysteine in 10% MMR (pH 8.0) for TETS.

ChIP-sequencing

Embryos (n = 35–90, two biological replicates for every ChIP experiment) were fixed in 1% formaldehyde for 30 min at Nieuwkoop-Faber stage 10.5. Embryos were washed once in 125 mM glycine/25% MMR and twice in 25% MMR, homogenized on ice in sonication buffer (20 mM Tris•HCl, pH 8/10 mM KCl/1 mM EDTA/10% glycerol/5 mM DTT/0.125% Nonidet P-40, and protease inhibitor cocktail [Roche]). Homogenized embryos were sonicated for 20 min using a Bioruptor sonicator (Diagenode). Sonicated extract was centrifuged at top speed in a cold table-top centrifuge and supernatants (ChIP extracts) were snap frozen in liquid nitrogen and stored at − 20 °C until use. Before assembling the ChIP reaction, the ChIP extract was diluted with IP buffer (50 mM Tris•HCl, pH 8/100 mM NaCl/2 mM EDTA/1 mM DTT/1% Nonidet P-40, and protease inhibitor cocktail) and then incubated with 1–5 μg of antibody and 12.5 μL Prot A/G beads (Santa Cruz) for an overnight binding reaction on the rotating wheel in the cold room. The following antibodies were used: H3K4me3 (Abcam ab8580), H3K4me1 (Abcam ab8895), p300 (C-20, Santa Cruz sc-585), H3K36me3 (Abcam ab9050), and RNA polymerase II (Diagenode C15200004). The beads were sequentially washed, first with ChIP1 buffer (IP buffer plus 0.1% sodium deoxycholate), then ChIP2 buffer (ChIP1 buffer with 500 mM NaCl final concentration), then ChIP3 buffer (ChIP1 buffer with 250 mM LiCl), then again with ChIP1 buffer, and lastly with TE buffer (10 mM Tris, pH 8/1 mM EDTA). The material was eluted in 1% SDS in 0.1 M sodium bicarbonate. Cross-linking was reversed by adding 16 μL of 5 M NaCl and incubating at 65 °C for 4–5 h. DNA was extracted using the Qiagen QIAquick PCR purification kit. Approximately 10 ng input DNA was used for sample preparation for high-throughput sequencing on an Illumina HiSeq 2000 or NextSeq (according to manufacturer's protocol).

RNA-sequencing

For RNA-seq experiments, total RNA was extracted from 20 Nieuwkoop-Faber stage 10.5 embryos (two biological replicates each for LELS and LETS, respectively) using Trizol and Qiagen columns. In total, 4–5 μg total RNA was treated with DNase I on column and depleted of ribosomal RNA (rRNA) using Magnetic gold RiboZero RNA kit (Illumina) resulting in a yield of 45–52 ng of rRNA depleted total RNA. A total of 2 ng rRNA-depleted total RNA was reserved for Experion (Bio-Rad) quality assessment run for rRNA depletion and the remaining was used for first and second strand synthesis (strand-specific protocol). Total yield of double-strand DNA (dscDNA) was in the range of 14.5–15.8 ng and out of this 1.2–5 ng was used for sample preparation for high high-throughput sequencing (according to manufacturer's protocol). Quantitative polymerase chain reaction quality controls before and after sample preparation corroborated well and relative depletion of 28S rRNA compared to control genes (eef1a1 and gs17) was taken as a quality assessment indicator for sequencing-grade dscDNA.

ChIP-seq and RNA-seq data analysis

ChIP-seq reads were mapped to the *X. laevis* genome (Xenla9.1) using bwa mem (version 0.7.10-r789) with default settings [36]. Duplicate reads were marked using bamUtil v1.0.2. Where applicable (H3K4me3, p300) peaks were called using MACS (version 2.1.0.20140616) [37] relative to the Input track using the options –broad -g 2.3e9 -q 0.001. –buffer-size 1000. Peaks were combined for replicates using bedtools intersect (version v.2.20.1) [38]. Figures of genomic profiles were generated using fluff v1.62 [39].

In addition to the RNA-seq triplicate produced in this study, we used the eight stage 10.5 samples from NCBI

GEO series GSE56586 (GSM1430926, GSM1430927, GSM1430928, GSM1430929, GSM1430930, GSM143 0931, GSM1430932, GSM1430933). RNA-seq reads were mapped to the Xenla9.1 genome with the JGI 1.8 annotation using STAR version 2.4.2a [40]. Quantification of expression levels was performed using express eXpress version 1.5.1 [41]. The mean expression level (TPM; transcript per million) per transcript was obtained by combining all replicates.

MethylC-seq for whole-genome bisulfite sequencing

Genomic DNA from *Xenopus* embryos (LELS and LETS, n = 20–50, NF stage 10.5) was extracted as described before [42] with minor modifications. Briefly, embryos were homogenized in 3 volumes STOP-buffer (15 mM EDTA, 10 mM Tris-HCl pH7.5, 1% SDS, 0.5 mg/mL proteinase K). The homogenate was incubated for 4 h at 37 °C. Two phenol:chloroform:isoamyl alcohol (PCI, 25:24:1) extractions were performed by adding 1 volume of PCI, rotating for 30 min at room temperature (RT) and spinning for 5 min at 13 k rpm. DNA was precipitated in 1/5 volume NH4AC 4 M plus 3 volumes EtOH with an overnight incubation at 4 °C. Subsequently, the DNA was spun down for 20 min at 13 k rpm in a cold centrifuge and the pellet was washed with 70% EtOH and dissolved in 100 μL of DNAse-free water. To remove contaminating RNA, a 2-h RNase A (0.01 volume of 10 mg/mL) treatment was performed at 37 °C. Sample was further purified with two Mg/SDS precipitations. Volumes of 0.05 of 10% SDS plus 0.042 volumes of MgCl2 2 M were added to the sample followed by incubation on ice for 15 min. Subsequently, the precipitants were spun down at 4 °C for 5 min at 13 k rpm. A third PCI extraction was also performed followed by only one chloroform:isoamyl alcohol (CI, 24:1) extraction. DNA was precipitated overnight at − 20 °C in 2.5 volumes EtOH plus 1/10 volume NaOAc 3 M pH 5.2. Next, the precipitated DNA was spun down for 30 min at 13 k rpm in a cold centrifuge and the pellet was washed with 70% EtOH. The purified DNA pellet was then dissolved in 50 μL H2O.

MethylC-seq library generation was performed as described previously [43, 44]. The genomic DNA was sonicated to an average size of 200 bp, purified and end-repaired followed by the ligation of methylated Illumina TruSeq sequencing adapters. Library amplification was performed with KAPA HiFi HotStart Uracil + DNA polymerase (Kapa Biosystems, Woburn, MA, USA), using six cycles of amplification. MethylC-seq libraries were sequenced in single-end mode on the Illumina HiSeq 1500 platform. The sequenced reads in FASTQ format were mapped to the in-silico bisulfite-converted *X. laevis* reference genome (Xenla9.1) using the Bowtie alignment algorithm with the following parameters: -e 120 -l 20 -n 0 as previously reported [45, 46]. DMRs were called using the methylpy pipeline, as described before [46], with FDR < 0.05 and the difference in fraction methylated ≥ 0.4. To estimate the bisulfite non-conversion frequency, the frequency of all cytosine base-calls at reference cytosine positions in the lambda genome (unmethylated spike in control) was normalized by the total number of base-calls at reference cytosine positions in the lambda genome. See below for sequencing and conversion statistics.

DNA-methylation free (hypo-methylated) regions were detected using the hmr tool from MethPipe version 3.0.0 (http://smithlabresearch.org/software/methpipe/) [47]

Active transcription

To consider a region as actively transcribed, we measured the H3K36me3 and RNAPII marks (as RPKM) of 200,000 random regions in *X. laevis* to define background levels. Regions with active transcription are those with at least the average of the measures plus two standard deviations, for both signals independently.

Whole-genome alignment

Genome alignment of *X. tropicalis* and *X. laevis* was performed using progressiveCactus version 0.0 (https://github.com/glennhickey/progressiveCactus) [39, 40] with the default parameters. *X. tropicalis. X. laevis* L and S were treated as separate genomes and were aligned using (Xla.v91.L:0.2,Xla.v91.S:0.2):0.4,xt9:0.6) Newick format phylogenetic tree. In order to reduce computational time alignment was done per-chromosome, with homeologous chromosomes aligned to each other.

Calling deletions

A set of high-confidence deleted regions was obtained using the progressiveCactus alignment. We extracted all regions from the *X. laevis* genome that reciprocally aligned either *X. tropicalis* and/or to the other subgenome. We then selected all regions that reciprocally aligned to *X. tropicalis* but not to the other *X. laevis* subgenome. We merged all regions within 10 bp and removed those that overlapped for > 25% of their length with gaps. As a final filtering step, we required a sequence that reciprocally aligned to the other subgenome in both 500-bp flanks of the putative deletion. Finally, the size of the region between the two aligned flanks should be at most 4 kb and at least three times shorter than the size of the region in the subgenome where the sequence was not deleted.

SNP calling

SNPs were called using the GATK pipeline (version 3.4-46-gbc02625 [48]) on the basis of the best practices workflow [49, 50] As input we used a high-

coverage ChIP-input track from a clutch of wild-type embryos compared the reference J-strain genome. The HaplotypeCaller tool was used to call SNPs. All putative SNPs were subsequently filtered with the VariantFiltration tool. The filterExpression was set to "QD < 2 || FS > 60.0 || MQ < 35.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" for *X. tropicalis*. For *X. laevis* the same settings were used, except for MQ, which was set to "MQ < 40." SNPs passing the filter were required to have at least tenfold coverage with at least four observations of the alternative allele. The SNP coverage was calculated relative to the sequence regions where SNPs could be called given the minimum required coverage, as determined by the CallableLoci tool from the GATK pipeline.

Search and alignment of orthologs and evolution rates

Orthologs of *X. tropicalis* were searched in the genome of *X. laevis* with the cdna2genome tool from Exonerate [51]. From 14,500 sequences submitted, 14,276 were successfully scanned. From those, 10,935 found a match in both subgenomes, leaving 3343 sequences that did not return any sequence from either L or S subgenomes or both. Among the sequences with a match in both subgenomes, those having no synteny (n = 939) were discarded because they were potential wrong matches in closely related gene families.

Once we had our three sequences per gene (n = 9996), we aligned them using MACSE [52], which allows frameshifts and premature stop codons, with the following parameters: gap creation = 18, gap extension = 8, frameshift creation = 28, premature stop codon = 50. Ten sequences were discarded in this step.

In order to obtain evolutionary rates of each of the three copies per gene triangle, we performed ancestral sequence reconstruction with FastML [53], which gave us the most likely sequence present at the speciation between *X. laevis* L and S ancestors. Once we obtained this crossroad sequence, we measured the amount of ratio of non-synonymous mutations per non-synonymous sites versus synonymous mutations per synonymous sites (i.e. Ka/Ks ratio) using the seqinR package [54].

Pseudogene dating

Similar to Zhang et al. [21], we related the excess of non-synonymous mutations to the evolving rate average of the gene to date the approximate time when the copy lost constraint on its sequence.

Bootstrapping pseudogene dates

We took the pseudogene candidates and retrieved their annotated 1 to 1 orthologs in human, mouse, and chicken through Ensembl. We then aligned them using MACSE [52] with default parameters, considering the pseudogene as a "less reliable" sequence. After this, we reconstructed the ancestral sequence with FastML [53] and then measured the Ka/Ks ratio using the seqinR package [54].

In order to confirm the reliability of these results, we bootstrapped the alignments 1000 times each and measured the Ka/Ks ratios of all of them. Briefly, we cut up the alignments in codons and we built an artificial alignment of the same length of the original protein by randomly adding (with replacement) aligned codons found in the original alignment.

Quantification of genomic losses per genomic region

Using the deletions track generated through the deletions call step (see "Calling deletions" in "Methods"), we quantified the amount of DNA lost per genomic region by measuring the overlap between both coordinates. To do so, we used the R packages rtracklayer [55] and GenomicRanges [56]. To compare the observed distribution of deletions to the expected distribution, we performed 1000 genomic randomizations of the deletions, keeping features on the same chromosome, using bedtools shuffle [38] with the -chrom argument. *P* values for enrichment or depletion of overlap with specific features were calculated based on the z-score obtained from the 1000 randomizations. *P* values for differences in observed/expected rate between L and S chromosomes were calculated using the Mann–Whitney U test. All *P* values were adjusted for multiple testing using the Benjamini–Hochberg approach.

Gene Ontology term enrichment analysis

Term enrichment analysis was performed using PANTHER [57]. Briefly, we used *X. tropicalis* orthologs names of the pseudogenes discussed in the section "High levels of pseudogenization started after hybridization and continue to the present" and we compared it to the list of genes in *X. tropicalis* that successfully returned syntenic orthologs in *X. laevis* (see "Search and alignment of orthologs and evolution rates" in "Methods").

Quantification of preferential loss of complete protein complexes

We took the hetero-dimers from the human protein complex CORUM database [58] and examined the extent to, when completely represented in the *X. laevis* genome (357 complexes), both genes were present on both genomes (170 complexes), only one gene was present on both genomes (124 complexes) or both genes were present on only a single genome (63 complexes). Also, extending the analysis to trimers did not show an overrepresentation of completely lost complexes.

Additional files

Additional file 1: Overview of the sequencing data used in this study. (XLS 16 kb)

Additional file 2: Containing the number of experimentally defined regulatory regions (p300, H3K4me3, and DNAme-hypomethylated) along with the genomic coordinates of all of them. (XLS 5598 kb)

Additional file 3: Containing supplemental figures. (DOCX 1833 kb)

Additional file 4: Summarizing the analysis of small nucleotide polymorphisms (SNP) in different regions of the genome. (XLS 62 kb)

Additional file 5: With a supplemental note on the time required to fix pre-existing genomic variation in the population after hybridization. (DOC 38 kb)

Acknowledgements
The authors thank Ulrike J. Jacobi and Kees-Jan François for valuable contributions in an early phase of the work, Emese Gazdag for genomic DNA, and Shamil Sunyaev for helpful discussions.

Funding
This work has been supported by the US National Institutes of Health (NICHD, grant R01HD069344). Part of this work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation. DME and MAH were supported by the Virgo consortium, funded by the Dutch government (FES0908). RG was supported by an HFSP long-term fellowship LT 0004252014-L. RH was supported by R35 GM118183. SJvH is supported by the Netherlands Organization for Scientific research (NWO-ALW, grant 863.12.002). OB is supported by an Australian Research Council Discovery Early Career Researcher Award - DECRA (DE140101962).

Availability of data and materials
The data have been deposited in NCBI's Gene Expression Omnibus [35] and are accessible through GEO Series accession numbers GSE76059 (*X. laevis* ChIP-seq) [59], GSE92382 (genomic DNA; *X. laevis* RNA-seq; *X. tropicalis* × *X. laevis* ChIP-seq) [60], GSE90898 (*X. tropicalis* × *X. laevis* WGBS) [61], and GSE67974 (*X. tropicalis* ChIP-seq) [62].

Authors' contributions
ChIP-seq, RNA-seq data generation, and experimental design was performed by SSP with help from IvK, RG, and RH. GJCV, SJvH, and MAH designed the study. DME and GG were involved in analysis design. Bisulfite sample generation and sequencing was done by SSP, IvK, OB, and RL. OB and GG performed analysis of differentially methylated regions. Genome alignment and hybrid analysis was performed by GG. Analysis of deleted regions and SNPs was performed by SJvH and DME. DME also performed analysis of mutation rates and pseudogenes. DME, SSP, GG, MAH, SJvH, and GJCV wrote the paper. DME, SSP, and GG contributed equally to the study. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Radboud University Medical Center, Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands. [2]Radboud University, Faculty of Science, Department of Molecular Developmental Biology, Radboud Institute for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands. [3]Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, Australia. [4]St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, Australia. [5]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, Australia. [6]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [7]Harry Perkins Institute of Medical Research and ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Perth, WA 6009, Australia.

Received: 6 April 2017 Accepted: 3 October 2017
Published online: 24 October 2017

References
1. Wendel JF. The wondrous cycles of polyploidy in plants. Am J Bot. 2015;102:1753–6.
2. Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 2016;30:159–65.
3. Ohno S. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. Semin Cell Dev Biol. 1999;10:517–22.
4. Holland PW, Garcia-Fernandez J. Hox genes and chordate evolution. Dev Biol. 1996;173:382–95.
5. Grant SG. The molecular evolution of the vertebrate behavioural repertoire. Philos Trans R Soc Lond B Biol Sci. 2016;371:20150051.
6. Lien S, Koop BF, Sandve SR, Miller JR, Kent MP, Nome T, et al. The Atlantic salmon genome provides insights into rediploidization. Nature. 2016;533:200–5.
7. Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC. A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. Mol Phylogenet Evol. 2004;33:197–213.
8. Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, et al. Genome evolution in the allotetraploid frog Xenopus laevis. Nature. 2016; 538:336–43.
9. Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M. Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol. 2014;31:448–54.
10. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 2011;108:4069–74.
11. Song Q, Chen ZJ. Epigenetic and developmental regulation in plant polyploids. Curr Opin Plant Biol. 2015;24:101–9.
12. Bhaumik SR, Smith E, Shilatifard A. Covalent modifications of histones during development and disease pathogenesis. Nat Struct Mol Biol. 2007;14:1008–16.
13. Perino M, Veenstra GJ. Chromatin control of developmental dynamics and plasticity. Dev Cell. 2016;38:610–20.
14. Carroll SB. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell. 2008;134:25–36.
15. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
16. Hontelez S, van Kruijsbergen I, Georgiou G, van Heeringen SJ, Bogdanovic O, Lister R, et al. Embryonic transcription is controlled by maternally defined chromatin state. Nat Commun. 2015;6:10148.
17. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015;160:554–66.
18. Matsuda Y, Uno Y, Kondo M, Gilchrist MJ, Zorn AM, Rokhsar DS, et al. A new nomenclature of Xenopus laevis chromosomes based on the phylogenetic relationship to Silurana/Xenopus tropicalis. Cytogenet Genome Res. 2015; 145:187–91.
19. Weckselblatt B, Rudd MK. Human structural variation: mechanisms of chromosome rearrangements. Trends Genet. 2015;31:587–99.
20. Subramanian S, Kumar S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 2003;13:838–44.
21. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol. 2010;11:R26.
22. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic Acids Res. 2015;43:e101.
23. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 2013;5:28.

24. Narbonne P, Simpson DE, Gurdon JB. Deficient induction response in a Xenopus nucleocytoplasmic hybrid. PLoS Biol. 2011;9:e1001197.
25. van Kruijsbergen I, Hontelez S, Elurbe DM, van Heeringen SJ, Huynen MA, Veenstra GJ. Heterochromatic histone modifications at transposons in Xenopus tropicalis embryos. Dev Biol. 2017;426:460–71.
26. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8:272–85.
27. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. Human mutation rate associated with DNA replication timing. Nat Genet. 2009;41:393–5.
28. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134:1289–303.
29. Sasaki M, Lange J, Keeney S. Genome destabilization by homologous recombination in the germ line. Nat Rev Mol Cell Biol. 2010;11:182–95.
30. Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, et al. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature. 2016;530:171–6.
31. Patel A, Horton JR, Wilson GG, Zhang X, Cheng X. Structural basis for human PRDM9 action at recombination hot spots. Genes Dev. 2016;30:257–65.
32. Nishihara H, Kobayashi N, Kimura-Yoshida C, Yan K, Bormuth O, Ding Q, et al. Coordinately co-opted multiple transposable elements constitute an enhancer for wnt5a expression in the mammalian secondary palate. PLoS Genet. 2016;12:e1006380.
33. de Souza FS, Franchini LF, Rubinstein M. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol. 2013;30:1239–51.
34. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24:1963–76.
35. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet. 2014;15:221–33.
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
37. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
38. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
39. Georgiou G, van Heeringen SJ. fluff: exploratory analysis and visualization of high-throughput sequencing data. Peer J. 2016;4:e2209.
40. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
41. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods. 2013;10:71–3.
42. Bogdanovic O, Fernandez-Minan A, Tena JJ, de la Calle-Mustienes E, Gomez-Skarmeta JL. The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. Methods. 2013;62:207–15.
43. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature. 2011;471:68–73.
44. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc. 2015;10:475–83.
45. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
46. Bogdanovic O, Smits AH, de la Calle ME, Tena JJ, Ford E, Williams R, et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat Genet. 2016;48:417–26.
47. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One. 2013;8:e81148.
48. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
49. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
50. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11 10 11-33.
51. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
52. Ranwez V, Harispe S, Delsuc F, Douzery EJ. MACSE: multiple alignment of coding SEquences accounting for frameshifts and stop codons. PLoS One. 2011;6:e22594.
53. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012;40:W580–584.
54. Charif D, Lobry JR. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural approaches to sequence evolution: molecules, networks, populations. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 207–32.
55. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–2.
56. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9:e1003118.
57. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8:1551–66.
58. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. 2010;38:D497–501.
59. van Heeringen SJ, Paranjpe SS, Veenstra GJC. ChIP-sequencing in stage 10.5 Xenopus laevis embryos. Gene Expression Omnibus. 2016, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76059.
60. Paranjpe SS, Georgiou G, van Kruijsbergen I, Gibeaux R, Heald R, van Heeringen SJ, et al. Regulatory remodeling in the allo-tetraploid frog Xenopus laevis. Gene Expression Omnibus. 2017, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92382.
61. Bogdanovic O, Lister R. Single-base resolution methylomes of Xenopus laevis x Xenopus tropicalis embryos. Gene Expression Omnibus. 2017, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE90898.
62. Hontelez S, Veenstra GJC. Embryonic transcription is controlled by maternally defined chromatin state. Gene Expression Omnibus. 2015, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67974.